Plan Overview

A Data Management Plan created using DMPTool-Stage

Title: Comportamiento de la movilidad en bicicleta considerando las condiciones climáticas y los puntos turísticos en la ciudad de New York.

Creator: Carlos Sánchez

Affiliation: Universidad de Los Andes (uniandes.edu.co)

Principal Investigator: Federico Higuera

Data Manager: Cristian Ospina

Project Administrator: Carlos Sanchez

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Start date: 10-23-2022

End date: 01-02-2023

Last modified: 11-06-2022

Copyright information:

Comportamiento de la movilidad en bicicleta considerando las condiciones climáticas y los puntos turísticos en la ciudad de New York.

En este proyecto se va a recibir datos de **viajes en bicicleta**, **estaciones del clima** con información diaria y **puntos de interés**, todo lo anterior dentro de la ciudad de Nueva York.

El dataset de **viajes en bicicleta** tiene 135 millones de registros y contiene información como la duración del viaje, el punto de inicio y fin, el género de la persona que uso la bicicleta y el tipo de bicicleta. El dataset de la **información del clima** tiene 229.327 registros mientras que la base de **puntos de interés** cuenta con 20.568 registros. Todos los datasets están en **formato csv**, esto permite el fácil acceso, mantenimiento y divulgación de la información.

La información es recolectada a partir de archivos en formato csv los cuales han sido suministrados a los investigadores del proyecto.

La primera base de datos contiene información de los **viajes en bicicleta** y contiene 92 archivos comprimidos, cada archivo representa los viajes correspondientes a 1 mes en los periodos comprendidos entre junio 2013 y octubre 2021. La nomenclatura seguida para el almacenamiento de estos archivos es:

<Año><Mes>-citybike-tripdata.csv.gz

Donde los campos Año y Mes son remplazados por el periodo correspondiente. Adicionalmente, todos estos archivos están almacenados dentro de la carpeta citybikegz.

La información del clima viene comprimida, en 3 archivos cuya nomenclatura es la siguiente.

Weather_<Año Inicial>-<Año Final>.csv

Estos archivos están almacenados dentro de la carpeta clima.

El primero de ellos contiene datos desde el año 2013 al 2015

El segundo archivo desde el 2016 al 2018

Finalmente, el tercer archivo contiene datos del 2019 al 2021

Los datos de los puntos de interés de la ciudad vienen en un solo archivo con toda la información.

Este es llamado point_of_interest.csv

Para la primera base de datos, ‘citibikegz’, no se tiene una documentación formal definida, sin embargo, su interpretación es intuitiva a partir de los nombres de las columnas que conforman el dataset, dado que está explicado su contenido en Inglés.
Para el caso del dataset ‘clima’ se cuenta con un archivo en formato .pdf, que expone a detalle la información contenida en el dataset, así como explica la interpretación de los valores encontrados en las diferentes columnas del mismo, adicionalmente, se presenta un enlace en donde se puede encontrar información complementaria sobre el dataset en cuestión. Dicha documentación se incluye dentro de la carpeta que contiene los archivos que componen la base de datos.

El dataset de ‘Point_of_interest’ cuenta también con un archivo en formato .pdf que presenta a detalle la información que se presenta en la base de datos, especificando el significado de cada uno de los campos y su interpretación. También se cuenta con la metadata del mismo. Dicha información se encuentra contenida dentro de la carpeta donde se almacena el archivo de datos.

En general, la información que se maneja no corresponde a datos personales ni información sensible, por lo tanto, no habría consideraciones éticas adicionales que hacer. Además, todos los análisis planeados en el proyecto tienen un propósito benigno y no incurre en temas de privacidad personal.

En el caso de que el proyecto incurra en el manejo de información personal, se debe solicitar de manera explícita el consentimiento para conservar y compartir los datos a las personas involucradas.

La información producida por este equipo de investigación será de dominio público y tiene como objetivo ayudar a la comunidad. Por lo tanto, en términos de copyright y propiedad intelectual solo es necesario otorgar el reconocimiento necesario según la ley.

Se tiene planeado la publicación de un repositorio público en la plataforma Github con el objetivo de presentar los resultados y conclusiones, mientras que se asegura la transparencia de la investigación.

La información correspondiente será almacenada en una bodega de datos en la nube provista por la institución universitaria, en este caso, el departamento de Ingeniería de Sistemas de la Universidad de los Andes. Para comunicarse con la base de datos en la nube se utiliza Google Cloud como proveedor de servicios.

Los datos se encuentran en una instancia de Google Cloud a la cual únicamente es posible acceder a través de las credenciales dadas por la Universidad, quien es la responsable de administrar el acceso y la colaboración de los integrantes del proyecto. El acceso a la información será libre e ilimitado para los 3 investigadores jefes con el propósito de no retrasar la investigación. A pesar de esto, la información utilizada para los análisis es de carácter público, por lo tanto, cualquier persona podría acceder a esta de forma libre vía internet.

Se considera relevante para su almacenamiento y publicación los datos obtenidos luego de la realización del proceso de ETL en el cual se ejecutó la compresión de los datos contenidos en los 3 datasets para su unificación en una única base de datos.

A partir de estos datos será posible la realización de diferentes análisis que involucren los factores de desplazamiento, periodo de tiempo, condiciones metereológicas y visitas de puntos de interés en la ciudad de New York.

Actualmente, los datos están almacenados en la instancia de Google Cloud, la cual tiene un costo mensual por mantener esa información.

El procedimiento y los resultados obtenidos serán publicados en GitHub. Este servicio es gratuito y, por lo tanto, no genera costo para el proyecto.

Los datos originales necesarios para el desarrollo del proyecto únicamente serán accesibles para el equipo investigador. Sin embargo, los resultados y las conclusiones podrán ser encontradas en el repositorio público de GitHub para lo cual los usuarios potenciales deben realizar la solicitud de pull de los mismos.

Como se mencionó anteriormente, la información solo está disponible para el equipo investigador durante la duración del proyecto. Posteriormente, será de uso público por medio de la solicitud en a partir de un pull al

repositorio de Github.

El gobierno de los datos será responsabilidad del equipo investigador.

Para ejecutar el plan de desarrollo del proyecto es necesario capacidad de almacenamiento y poder computacional. Ambas necesidades serán cubiertas por la Universidad de los Andes a partir de la generación de instancias virtuales con el proveedor Google Cloud.

Se requiere soporte por parte del equipo de la Universidad para resolver cualquier duda respecto al manejo y la configuración de estos recursos.
